Description

FOWLER-NORDHEIM BLOCK ALTERABLE
EEPROM MEMORY CELL

5

TECHNICAL FIELD

The present invention relates in general to
semiconductor devices. More specifically, the present
invention relates to block alterable memory devices.

10

BACKGROUND ART

The need for a high-density block alterable
memory devices is ever increasing. Cellular phones,
memory sticks, digital cameras, laptop computers, and
15   personal data assistants are a few examples of small
devices that demand higher density memories. These
devices require alterable memories because their contents
change every time they are in use. For example, the size
of a memory stick is as small as a pen but it can store
20   256 MB memory. The memory stick has a Universal Standard
Bus (USB) port that can plug into another USB memory port
of a computer to transfer the data from the hard drive of
that computer. Therefore, the memory stick and other
similar devices such as camera memories need a high-
25   density alterable memory device to erase old data and
store new data. The Electrically erasable programmable
read only memory (EEPROM) common in the industry cannot
be used in these applications because EEPROM is not
alterable under normal operation conditions.

30   A typical block alterable memory device employs
flash memory to program, read, or erase memory cells.
With reference to Fig. 1, a flash memory 100 is a memory
array which is arranged in rows 102 and columns 106.
Each row 102 has N+1 memory cells connecting to source
35   lines S_0 – S_N. The first memory cell in the row 102
belongs to column BL_0 and the Nth memory cell belongs

column BL_N. Therefore, there are N+1 columns in the flash memory array 100. The gates of all the cells within a column 106 are coupled together to form a wordline WL_i 102. There are M+1 wordlines or rows in the flash memory array 100, ranging from WL_0 to WL_M. The sources of the cells in each column are coupled together and coupled to the select lines 104, ranging from S_0 to S_N. The drains of the cells in each row are coupled together to form a bitline 106, ranging from BL_0 to BL_N. The flash array 100 enables users to electrically program and erase information stored in a memory cell 110.

Each memory cell 110 in the flash memory matrix 100 is a floating gate transistor. The structure of a floating gate transistor is similar to a traditional MOS device, except that an extra polysilicon strip is inserted between the gate and the channel. This strip is not connected to anything and called a floating gate. The threshold voltage of a floating gate transistor is programmable. The described flash memory 100 uses the Fowler-Nordheim tunneling effect to program a cell 110. Programming is a process wherein electrons are placed in the floating gate. Programming occurs when applying a high voltage between the gate and source, and gate and drain terminals that a high electric field causes injection of carriers into the floating gate. Electrons acquire sufficient energy to become hot and traverse the first oxide insulator, so they get trapped on the floating gate. Programming is done on a bit basis by applying a correct voltage at the bitline 106 of each cell 110.

The floating gate layer allows the cell 110 to be electrically erased through the gate. Erase and program operations of the memory array 100 can be done on more than one cell at a time. However, the alterable

flash memory device has reliability and durability problems because the voltages for erasing and programming are very high.

With reference to Table 1 at the end of this specification, in order to achieve block alterable memory, the memory cell 110 in the flash memory array 100 as shown in Fig. 1 needs to apply +10 volts or -10 volts across the wordline WL_i 102, the source line S_i 104, and the bitline BL_i 106. Accordingly, the placement of such high voltages to a single memory cell transistor 110 presents reliability and durability problems. Over long periods of time, placing high voltages on the memory device 100 may alter a program stored in each cell 110.

One prior art solution to this problem (for example, U.S. Patent No. 5,066,992 to T.C. Wu) is shown in Fig. 2A. This solution places an extra select transistor 202A in series with a flash memory cell 210A. The gate of the additional select transistor 202A is coupled to the select line S_0 to S_N, the drain 204A is coupled to the bitline BL_0 to BL_N 214A, and the source 206A is coupled to the drain of the flash cell. Thus, when a select line S_i is ON, each selected transistor connected to the select line S_i is turned ON. As a result, the voltage of the drain of the flash cell 210A is proportional to the voltage of the bitline B_i. During a reading cycle, the bitline 214A is open, the select line S_i is grounded, and the wordline WL_i is at negative program voltage V_D. Thus, a program stored in an EEPROM device 200A remains unaltered. Thus, the memory array 100 lasts longer and avoids the reliability and durability of one-transistor memory cells presented above. However, the two-transistor memory cells require larger areas for manufacturing because each memory cell has two transistors.

Referring to Fig. 2B, various cross-sectional views of a memory array 200B are shown. Memory array

200B is formed on a face of a semiconductor substrate
222B. Substrate 222B is doped with a p-type majority
carrier. Bitline BL 214B, select line SEL 202B, wordline
WL 208B, and the source are n-type and implanted within
5   substrate 222B at the surface. The gate 208B comprises a
first poly layer 209B, a second poly layer 211B, and an
inter poly layer 212B. Accordingly, column lines 214B
and 206B serves as a source and drain of transistors
which are used in forming memory cells contained within
10   memory array 200B. Each of the column lines 214B serves
as a source of one memory cell or a drain of an adjacent
cell. However, this solution dedicates large sections on
the semiconductor substrate to the alterable block
function. An undesirably low density flash memory
15   results. Consequently, the industry has a need for a
memory device structure which has block alterable
capability without dedicating semiconductor substrate
area to that function.

       U.S. Patent No. 4,783,766 to Samachisa et al.
20   describes a memory cell of a block alterable EEPROM in
which a single control gate is common to both the
floating gate memory cell and the select transistor
device. However, the device is formed using a different
process flow from that of flash memory devices, thus
25   requiring a separate masking sequence.

       U.S. Patent No. 6,420,753 to Hoang describes a
similar structure to that of the Samachisa patent. It is
stated that these memory cells can be manufactured
without requiring additional processing steps from those
30   of comparable flash memories.

SUMMARY OF THE INVENTION
       A Fowler-Nordheim block alterable memory cell
in accordance with the present invention is carried out
35   in one form by a memory cell constructed from two
separate transistor cells that have common select-control

gate. The two cells are constructed on a substrate or in a well that exhibits a first (e.g., "p" or acceptor) conductivity type. A tunnel oxide layer resides on the substrate face. The select-control gate comprises a

5 first poly layer, an interpoly layer, and a second poly layer. The second poly layer is extended to connect to the gate of the first cell to form a common select-control layer. The extended portion of the common select-control layer contacts a drain implant region. A

10 buried n+ implant region is formed near the surface of the p-substrate. The floating gate region is positioned above the buried implant and extends over the channel of transistor 400B. A self-aligned source/drain implant is located at edges of the control poly. The area of the

15 substrate between the floating gate region and the drain implant region that lies underneath the extended portion of the common select-control layer is known as the active region. Thus, the Fowler-Nordheim block alterable memory device in accordance with the present invention is

20 constructed as a single transistor memory cell but it behaves as a two transistor cell because of the extended select-control layer.

   The present invention is also a method to fabricate a memory cell to achieve a Fowler-Nordheim

25 block alterable memory cell as described above. The method first deposits a screen oxide of about 150 angstroms thickness over the p-type substrate. Then a photoresist mask with an opening is added on top of the screen oxide layer. A cell channel and buried n⁺

30 implants are implanted at the location of the opening of the mask and near the surface of the p-substrate. Next, the screen oxide is etched and initial gate oxides are grown. A tunnel window mask is then formed. A tunnel oxide is etched in the screen oxide layer where the

35 windows of the tunnel window mask are located. The first polycrystalline silicon (poly) layer over the tunnel

oxide and cell implants are deposited. An insulating
layer is formed overlying the first poly layer. An
extended final (second) poly layer is deposited over the
insulating layer. Finally the device is completed by
5   source and drain implants.

BRIEF DESCRIPTION OF THE DRAWINGS
        Fig. 1 illustrates a schematic diagram of a
prior art memory array having single flash memory cells.
10   Fig. 2A illustrates a schematic diagram of a
prior art dual transistor memory cell that has block
alterable capability. The top cell is used to define the
block to be altered, and the second cell or flash cell is
used to store data information.
15   Fig. 2B illustrates a sectional view of the
dual transistor memory cell of Fig. 2A.
        Fig. 3 illustrates a schematic diagram of a
Fowler-Nordheim alterable block memory array in
accordance with the present invention.
20   Fig. 4A illustrates a schematic diagram of a
single cell from the Fowler-Nordheim alterable block
memory array in accordance with the present invention.
        Fig. 4B illustrates a cross sectional view of a
Fowler-Nordheim block alterable memory cell as
25   illustrated in Fig. 4A.
        Figs. 5A-5G illustrate exemplary cross
sectional views of process steps in accordance with the
present invention.
        Fig. 5A illustrates a cross sectional view of a
30   substrate with a screen oxide layer.
        Fig. 5B illustrates a cross sectional view of a
barrier mask layer with a window for depositing a cell
implant in the substrate.
        Fig. 5C illustrates a cross sectional view of a
35   mask-out implant.

Fig. 5D illustrates a cross sectional view of a p-substrate with a buried n+ implant, a source implant, and a drain implant.

Fig. 5E illustrates a cross sectional view of a Fowler-Nordheim cell with a window tunnel mask and etch oxide.

Fig. 5F illustrates a cross sectional view of a Fowler-Nordheim block alterable memory cell with a tunnel oxide layer and a first polycrystalline layer.

Fig. 5G illustrates cross sectional view of a Fowler-Nordheim block alterable cell with an oxide-nitride-oxide (ONO) deposition and a control poly layer deposition.

Fig. 6 illustrates a flowchart of a method for manufacturing the Fowler-Nordheim block alterable cell corresponding to Figs. 5A-5G.

BEST MODE FOR CARRYING OUT THE INVENTION

Figs. 3, 4A, and 4B show various views of an exemplary embodiment of a Fowler-Nordheim block alterable memory architecture fabricated according to a method of the present invention. With respect to Fig. 3, a memory array 300 includes a plurality of memory cells 310. The memory cell 310 includes a select transistor 302 and a memory transistor 304 which share a common control gate. The plurality of memory cells 310 may be erased and programmed in blocks or programmed or read bit-by-bit by applying appropriate voltages to the bitlines (BL_0 to BL_N), source lines (S_0 to S_N) and wordlines (WL_0 to WL_M). Typically all memory cells 310 in the array 300 are normally constructed as a result of similar process steps, and therefore, all cells are similar in structure.

Referring to Fig. 4A, a schematic of a memory cell 310 in accordance with the present invention includes a memory transistor 304 connected in series with a select transistor 302 at a drain/source junction. (The

drain of the memory transistor 304 is coupled to the source of the select transistor 302.) The source of the memory transistor 304 is coupled to a select line S_i. The drain of the select transistor 302 is coupled to a

5   bitline BL_i. A common gate of the memory transistor 304 and select transistor 302 is coupled to a wordline WL_j. This common gate for the memory 304 and select 302 transistors can be manufactured as a single cell having an extended and continuous poly layer, thus reducing an

10   area of the memory cell 310.

         With reference to Fig. 4B, a cross sectional view 400 of the memory cell 310 is illustrated. The memory cell 400 is formed on a semiconductor substrate (or well) 401 of a first conductivity type, which in the

15   exemplary embodiment is p-type. A drain implant region 402 and a source implant region 406, are implanted within an uppermost surface of the substrate 401. A buried, heavily doped implant 404 for the floating gate region is also formed within an uppermost surface of the substrate

20   401. The implant regions 402, 404, and 406 are of a second conductivity type of a polarity opposite that of the conductivity type of the substrate 401. In a specific exemplary embodiment, the implants are n-type. The buried implant 404 is of n+ conductivity and serves

25   as a tunneling charge source for a floating gate of the memory transistor 400. The drain implant region 402 and the buried implant 404 are spaced apart, so as to define an active region 414 therebetween. Accordingly, the drain implant region 402 connects to the bitline BL_i. The

30   source implant region 406 connects to the source line S_i.

         A first poly layer 410, forming the floating gate of the memory transistor 304, overlays the buried implant region 404, separated therefrom by a gate ONO layer 450. A second poly layer 408, forming a common

35   control gate, extends continuously over the first poly layer 410 (which forms the floating gate) from the source

dopant region 406 to the drain dopant region 402, overlaying both the buried implant 404 and the select transistor 302 active region 414. A tunnel oxide 460 of thickness 50-70 angstroms is formed in a tunnel window
5   region between the buried implant 404 and the floating gate 410B.

An exemplary manufacturing process of the memory cell 310 is shown in the flowchart of Fig. 6 and a result after each step is shown in Figs. 5A-5G. With
10   reference to Fig. 5A, according to a preferred process of manufacturing the present invention, at step 602, a screen oxide 504 is deposited over a substrate 502. The thickness of the screen oxide layer is approximately 150 angstroms.

15   Referring to Figs. 5B and 6, at step 604, a photoresist mask 506 is applied at face 504 of substrate 502. This mask 506 is patterned so as to permit ion implantation of a floating gate region though gaps in the photoresist mask 506. A buried N+ tunnel region 508 is
20   implanted in semiconductor substrate 502 through the opening of the mask 506 and the mask 506 is then removed using a conventional process. The substrate 502 is then annealed in, for example, a 900 °C nitrogen environment to ameliorate damage caused to substrate 502 by the prior
25   implantation step 604 and to diffuse the tunnel implant region 508 into substrate 502.

Referring to Figs. 5C and 6, at step 608, after the annealing treatment of the substrate 502, another mask 510 is formed on top of the oxide layer 504 for
30   memory cell implantation. Resulting cell implant regions 514 and 516 and buried implant region 512 are seen in Fig. 5D.

Referring to Figs. 5D and 6, at step 610, the screen oxide is etched away and an initial gate oxide
35   layer 517 is formed in its place.

Referring to Figs. 5E and 6, at step 612, a tunnel window mask 530 is deposited to a very high thickness so that an opening for a tunnel oxide layer 513 can be precisely positioned at the openings of this

5   tunnel window mask 530 layer above the buried implant 512.

With reference to Figs. 5F and 6, at step 614, after etching away the gate oxide layer 517 in the tunnel windows, a thin tunnel oxide layer 518 is deposited to a

10   thickness of about 50-70 angstroms. In a preferred embodiment the tunnel oxide layer 518 represents a thin, high quality silicon dioxide layer which may either be grown in a dry O_2 and HCl mixture atmosphere at a temperature of around 800° to 850 °C. Once the tunnel

15   oxide 518 has been formed, polysilicon floating gates 520 are formed over the gate 517 and tunnel oxide 518 layers.

Referring to Figs. 5G and 6, at step 616, an oxide or oxide nitride oxide (ONO) interpoly dielectric 521 is deposited and an etch is performed to create

20   interpoly insulation.

Next, a control gate poly layer (not shown) is applied using an LPCVD process. The deposition of the poly layer is a low temperature application, preferably at less than 625 °C, which tends to maintain the poly

25   layer in an amorphous state.

The poly layer is patterned and etched to produce strips of materials which form control gates 524. The control gate 524 polysilicon extends beyond the area above the floating gate 520 to adjacent areas to form a

30   common select gate. In addition, this pattern and etch step removes material from the poly layer thereby forming the remaining two sides for each of floating gates 520.

Finally, finishing step 618 is shown in Fig. 6, such as adding select transistor drain implants 528 and a

35   nitride overcoat may be performed to complete the process. Using the control gate 524 poly layer as a

mask, source implants 528 for the select transistor are made just past the outer edge of the control gate 524 poly.

A memory device constructed according to the
5 teaching of the present invention may be block erased and programmed, and also bit programmed. Referring to Table 1 and Fig. 3, in block programming, memory cell transistor sources, S_0 to S_N, in a block, and also the select transistor drains (the bitlines BL_0 to BL_N) are
10 held at a large negative potential, such as -10 volts, while the memory cell transistor control gates in the block (the wordlines WL_0 to WL_N) are raised to a relatively high positive voltage, such as 10 volts. This causes tunneling of electrons from the buried implant
15 through the tunnel oxide into the floating gates .

Memory cells may be block erased by leaving sources S_0 to S_N in the block open, and reversing the word and bitline voltages from the block programming case. Placing bitline electrodes in the block at a relatively
20 high positive voltage, such as 10 volts, and the wordline electrodes in the block at negative 10 volts, causes electrons be expelled out of the floating gate region back into the buried implant.

Bit programming involves applying a large
25 positive potential to the wordlines and to all bitlines except a selected bitline BL_{i+1}, which is at ground potential. The source lines S_0 to S_N are left open.

Memory cells in the present invention may be read by placing the control gate WL_{i+1} of a particular
30 cell ($_{i+1}$) to be read at positive V_D, and at the same time, placing the drain (bitline) of the particular cell to be read at a relatively low (about 1 volt) voltage V_s. All source lines S_0 to S_N are grounded in read mode. Cells not in the selected word (row) and bit column have

negative V_D voltages applied to their wordlines and bitlines that are open.

	WL_i	WL_{i+1}	S_i	BL_i	S_{i+1}	BL_{i+1}	S_{i+2}	BL_{i+2}
Block programming	+10V	+10V	−10V	−10V	−10V	−10V	−10V	−10V
Block Erase	−10V	−10V	Open	+10V	Open	+10V	Open	+10V
Bit (i+1) program	+10V	+10V	Open	+10V	Open	0V	Open	+10V
Read (I+1)	−VD	VD	GND	Open	GND	Vs~1V	GND	Open

Table 1: Voltages Required for Block Programming/Erasing in a Block Alterable Memory.